

Large scale clustering in structural and evolutionary analysis of SARS-CoV-2 proteins

Mitić N.^{1*}, Pavlović-Lažetić G.¹, Beljanski M.², Malkov S.¹, Maljković M.¹, Stojanović B.¹, Veljković A.¹, Kapunac S.¹

¹ Faculty of Mathematics, University of Belgrade, Belgrade, Serbia

² Institute of General and Physical Chemistry, BioLab, Belgrade, Serbia

* nenad@matf.bg.ac.rs

Key words: SARS-CoV-2, Relative Synonymous Codon Usage, proteins, WHO Annotation, clustering

Motivation and Aim: In order to understand SARS-CoV-2 origin, evolution and interaction with host's cells, various aspects of viral genome structure and function are under investigation. Codon usage (CU) frequency of viral proteins as well as non-silent mutations are of special interest, since they may contribute to changing virus characteristics. Previous analyses have shown that rare codons often occur in large clusters within protein coding sequences. In the case of SARS-CoV-2, previous codon usage analyses show an antagonistic codon usage pattern (i.e., use of rare codons) reducing translation speed, but increasing its precision, and yielding accurate and correctly folded viral proteins [1]. At that end, clustering of protein sequences are investigated based on Relative Synonymous Codon Usage (RSCU) as well as edit distances of amino acid (AAc) sequences providing for both characterizing (identifying) specific protein groups (types) and temporal evolution of proteins within groups (types). The paper will present results of such analysis.

Materials, Methods and Algorithms: A dataset of 423425 complete isolate nucleotide sequences have been extracted from <https://www.ncbi.nlm.nih.gov/sars-cov-2> on August 25, 2021. After cleaning process, remains 347962 isolates with 225.934 unique (2.366.031 total) SARS-CoV-2 protein coding nucleotide sequences, as well as the corresponding AAc sequences. Consistency check has been performed between the two based on standard genetic code table (transl_table 1). For all the proteins (141926) for which world-health-organization (WHO) SARS-CoV-2 annotation exists, submission date and protein sequence metadata are supplied and RSCU has been calculated for measuring CU bias in different proteins of different protein classes. Then different algorithms (including TwoStep clustering in SPSS Modeler program [2], hierarchical clustering in Cluto [3] and Python Scikit-learn library) were applied for k-clustering proteins based on RSCU, for $k = 2, 40$. Proteins of the most heterogeneous protein type – Surface glycoprotein (S-protein) – have been further clustered based on RSCU, for each WHO label and year/month date. Furthermore, for all the protein sequences, edit distances of AAc sequences for each pair of proteins have been calculated. Then different algorithms (e.g., spectral clustering) were applied for clustering proteins in each protein group. *Results:* Fig. 1. presents S-protein 18-clustering based on RSCU and labeled by WHO annotations. Clustering results are quite correct with silhouette of 0.47. The figure presents all the groups with quantities higher 10 %. Clustering performed by the SPSS Modeler program. Fig. 2 (for Epsilon WHO label) is representative of a set of figures presenting specific WHO groups on year/month scale when S proteins are clustered by the SPSS modeler into 18 clusters.

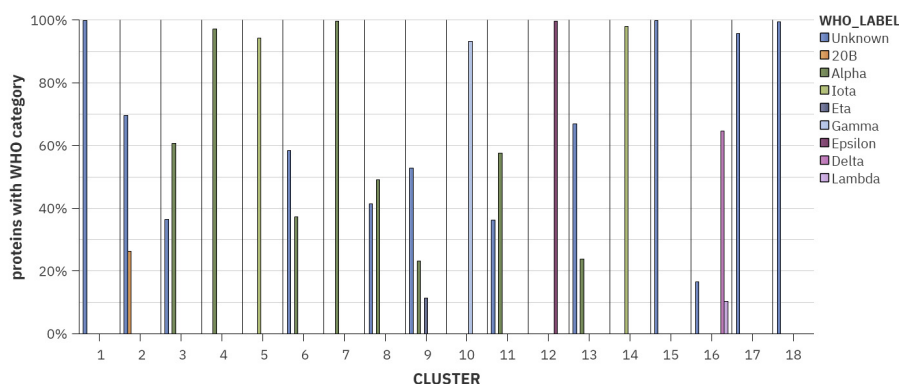


Fig. 1. S-protein 18-clustering based on RSCU and labeled by WHO annotations

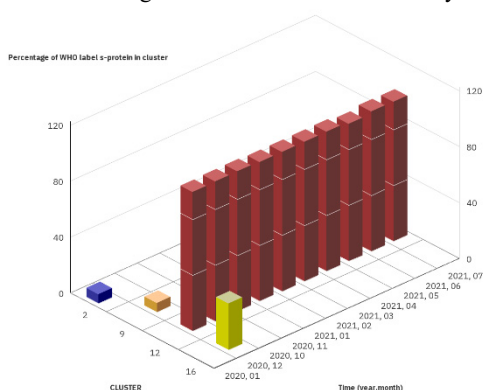


Fig. 2. Epsilon WHO group on year/month scale when S proteins are clustered by the SPSS modeler into 18 clusters

Fig. 2 is a representative of a set of figures presenting all the groups with quantities higher 5 %. Specific WHO groups mostly dominate in specific clusters in all the time periods (for example, as is the case with the Epsilon group dominating in the cluster 12). Spectral clustering of different types of proteins, based on AAc distances, give quite similar results regarding WHO labels, when applied to S protein (clusters rather homogenous), while less homogenous but still representative for other types of proteins. Hierarchical clustering of all the proteins for $k = 2,40$ produces highly homogenous clusters regarding protein types. Specifically, for $k = 12$ (the number of different protein types), each type is predominately represented by its specific cluster.

Conclusion: Since all the SARS-CoV-2 Orfs cluster in relatively homogenous clusters (according to WHO isolate classification), i.e., WHO-specifically annotated isolates make most of each cluster, this new approach may be used for annotation/prediction of strains that isolates belong.

References

1. Mogro E.G., Bottero D., Lozano M.J. Analysis of SARS-CoV-2 synonymous codon usage evolution throughout the COVID-19 pandemic. *Virology*. 2022;568:56-71.
2. IBM SPSS Modeler 18.2 Algorithms Guide, <https://www.ibm.com/support/pages/spss-modeler-182-documentation>
3. CLUTO – Software for Clustering High-Dim. Datasets: <http://glaros.dtc.umn.edu/gkhome/views/cluto>